Tracking planes in omnidirectional stereovision

Guillaume Caron, Eric Marchand and El Mustapha Mouaddib

Abstract Omnidirectional cameras allow direct tracking and motion estimation of planar regions in images during a long period of time. However, using only one camera leads to plane and trajectory reconstruction up to a scale factor. We propose to develop dense plane tracking based on omnidirectional stereovision to answer this issue. The presented method estimates simultaneously the parameters of several 3D planes along with the camera motion in a spherical model formulation. Results show the ef ciency of the approach.

I. Introduction

Motion estimation remains one of the key issue in the development of vision systems. Although the omnidirectional camera geometrical model (central projection model [3]) is now well known, few researches have been done to solve visual tracking problems. Omnidirectional cameras offer a very wide eld of view, allowing the perception of landmarks during a long period of time in image sequences, which is an important asset to the precision and robustness of an object tracking process.

Available tracking techniques uses either a 3D model leading mainly to a pose estimation process (eg, [8] in the perspective case and [14] for central catadioptric image) or 2D features, such as geometrical primitives (point, segments, circles, etc.) or object contours (such as active contours). One can also consider 2D models which can be a reference image (or a template). The goal is to estimate the motion (or warp) between the current image and the reference template. Examples of such approaches are differential tracking methods such as the KLT [1], [4], [9], [12].

In this paper, we consider a such approach within an omnidirectional stereovision sensor.

The warp to be estimated is a homography. This 2D motion model has the advantage to be exact for a plane whatever the camera motions are. This motion model was rst de ned for perspective camera model but is in fact valid for all central projection cameras [15], [16] of which some omnidirectional ones [2].

The central projection model have already been used for plane tracking in omnidirectional monocular images (eg [15]). This approach is similar to the one proposed in [4] although the registration between reference template and current image is done directly on the omnidirectional images. Nevertheless, the image gradients, needed to link image intensities to the variation of the motion, are computed in the image plane. However, it has been shown [6] that the ltering of omnidirectional images should be done on the spherical image rather than on the omnidirectional planar image. Indeed, the spherical geometry, for these images, is adapted to keep the properties of the convolution product used to compute gradients since it allows a regular sampling of the image whereas this is not possible in the image plane.

This work proposes to formulate the simultaneous tracking of image planar templates, constrained by homographies, directly within the spherical image space. Furthermore, rather than estimate a homography for each plane, we propose to estimate the camera pose, up to a scale factor, and recover the parameters of 3D planes. Based on this new spherical formulation, we then propose to use a stereoscopic catadioptric sensor allowing to perform the complete stereoscopic plane reconstruction, including scale factor, and tracking of spherical image planar regions. These three aspects together are the contribution of this work.

Thanks to the uni ed spherical projection model, the proposed stereoscopic plane reconstruction and tracking approach is generic and apply to all systems with multiple central cameras. But practically, stereoscopic systems may suffer from synchronization problems and even chromatic inconsistency between cameras. Hopefully, catadioptric omnidirectional stereovision can be achieved using a unique camera and several mirrors which solves the two latter problems. The used stereo catadioptric system is composed by an orthographic camera and four parabolic mirrors. This con guration (Fig. 1) has been proven to be optimal [7] and the image resolution in stereo tracking results is taken from this work.

The reminder of the paper presents the spherical tracking of a plane in monocular central vision, recalling the camera

(a) (b) (c)

Fig. 1. (a) Mirror part of a stereo sensor composed by (b) an orthographic camera and four coplanar paraboloidal mirrors [7] ; (c) the sensor on the mobile robot (reversed for stability reasons).

Guillaume Caron and El Mustapha Mouaddib are with Universit de Picardie Jules Verne, MIS laboratory, Amiens, FRANCE; e-mail: f guillaume.caron, mouaddibg@u-picardie.fr

Eric Marchand is with Universite de Rennes 1, IRISA, INRIA Lagadic, Rennes, France; e-mail: eric.marchand@irisa.fr

projection model and the projective transformation of a 3D plane motion before developing the spherical tracking itself. Then, the stereoscopic spherical plane tracking is presented. Finally, tracking results using a single camera and a stereo omnidirectional camera are presented, in static and mobile camera conditions. A conclusion highlighting the contributions of this work closes the paper.

II. Spherical plane tracking

A. Camera model

According to the uni ed model [3], a central projection camera can be modeled by a rst projection on a sphere, with a unit radius and a center $(0, 0, \xi)^\mathsf{T}$ expressed in the camera frame, followed by a perspective projection on the image plane. Such a model can be de ned using a unique parameter ξ which depends intrinsically on the central camera type.

Knowing intrinsic parameters $\gamma = \{p_x, p_y, u_0, v_0, \xi\}$, a 3D point $\mathbf{X} = (X, Y, Z)^\mathsf{T}$ is rst projected as $\mathbf{X}_\mathcal{S}$ on the unitary sphere:

$$\mathbf{X}_\mathcal{S} = (X_\mathcal{S}, Y_\mathcal{S}, Z_\mathcal{S})^\mathsf{T} = pr_\mathcal{S}(\mathbf{X}) = \left(\frac{X}{\rho}, \frac{Y}{\rho}, \frac{Z}{\rho}\right)^\mathsf{T} \quad (1)$$

where $\rho = \sqrt{X^2 + Y^2 + Z^2}$. Then this point is projected in the image plane as $\mathbf{x} = (x, y, 1)$. The relationship between $\mathbf{X}$ and $\mathbf{x}$ is expressed as:

$$\mathbf{x} = pr_\xi(\mathbf{X}) \quad \text{with} \quad x = \frac{X}{Z+\xi\rho} \quad \text{and} \quad y = \frac{Y}{Z+\xi\rho} \ . \quad (2)$$

$\mathbf{x}$ is the point on the normalized image plane and the image point in pixelic coordinates is obtained by the application of intrinsic parameters matrix $\mathbf{u} = \mathbf{K}\mathbf{x}$.

To back-project $\mathbf{x}$ to the equivalent sphere, the function $pr_\xi^{-1}(.)$ is used, knowing $\beta = \frac{\xi+\sqrt{1+(1-\xi^2)(x^2+y^2)}}{x^2+y^2+1}$: $\mathbf{X}_\mathcal{S} = pr_\xi^{-1}(\mathbf{x}) = (\beta x, \beta y, \beta - \xi)^\mathsf{T}$.

B. Projective transformation associated to a 3D plane

The coordinates of a point on the 3D plane in two views (for c_1 and c_2) are related by a homography. Considering the rotation ${}^{c_2}\mathbf{R}_{c_1}$ and the translation ${}^{c_2}\mathbf{t}_{c_1}$ between the two camera positions over time, the homography $\mathbf{H}$, de ned up to a scale factor [15], is given by:

$$\mathbf{H} \propto {}^{c_2}\mathbf{R}_{c_1} - \frac{{}^{c_2}\mathbf{t}_{c_1}\,{}^{c_1}\mathbf{n}^\mathsf{T}}{{}^{c_1}d}, \ {}^{c_2}\mathbf{t}_{c_1} \in \mathbb{R}^3 \text{ and } {}^{c_2}\mathbf{R}_{c_1} \in \mathbf{SO}(3) \quad (3)$$

where ${}^{c_1}\mathbf{n}$ is the normal to the plane of interest expressed in the rst camera frame and ${}^{c_1}d$ the orthogonal distance from the origin of this coordinates system to the plane. The homography is a 3×3 matrix and has 9 parameters, $h_{1,1}$ to $h_{3,3}$ and 8 degrees of freedom. To x the scale, $h_{3,3}$ is forced to be equal to 1.

According to the uni ed model, a central camera is modeled by a sphere and the homography $\mathbf{H}$ links two corresponding points ${}^{c_1}\mathbf{X}_\mathcal{S}$ and ${}^{c_2}\mathbf{X}_\mathcal{S}$ on two spheres, up to a scale factor:

$${}^{c_2}\mathbf{X}'_\mathcal{S} = \alpha\,{}^{c_2}\mathbf{X}_\mathcal{S} = \mathbf{H}\ {}^{c_1}\mathbf{X}_\mathcal{S}. \quad (4)$$

Actually, a homography relates two corresponding view lines and since we are working on the equivalent sphere, the spherical point ${}^{c_2}\mathbf{X}_\mathcal{S}$ corresponding to ${}^{c_1}\mathbf{X}_\mathcal{S}$, after being transfered using $\mathbf{H}$, is:

$${}^{c_2}\mathbf{X}_\mathcal{S} = {}^{c_2}\mathbf{X}'_\mathcal{S} \ / \ ||{}^{c_2}\mathbf{X}'_\mathcal{S}||. \quad (5)$$

C. Plane tracking on the unit sphere

To achieve plane tracking, we propose to use a differential minimization method that aims to compute the homography allowing to transfer spherical points of the reference sphere to the current one minimizing the sum of square differences of spherical image intensities. Even if tracking a plane directly estimating the coef cients of the projective transformation matrix $\mathbf{H}$ leads to the minimal set of parameters to estimate, the tracking is here de ned in the Euclidean space. Hence, the rotation, translation, plane normal and distance are explicitly estimated. Since the camera motion is the same for all the scene planes, this formulation allows to consider ef ciently multi planes tracking (see section II-D).

The criterion to minimize for the plane tracking is then:

$$\hat{\mathbf{H}} = \arg\min_{\mathbf{h}} \sum_{i=1}^{k} \left(\mathbf{I}_\mathcal{S}\left(\mathbf{H}\mathbf{X}^*_{\mathcal{S},i}\right) - \mathbf{I}_\mathcal{S}\left(\mathbf{X}^*_{\mathcal{S},i}\right)\right)^2 \quad (6)$$

where $\mathbf{X}^*_\mathcal{S}$ is a point of the reference region and the vector of parameters to estimate is $\mathbf{h} = (\mathbf{n}, d, \mathbf{r})^\mathsf{T}$. $\mathbf{r}$ is a vector representation of the relative pose and $\mathbf{H}$ is computed thanks to these parameters with the equation (3).

Considering a feature vector $\mathbf{I}_\mathcal{S}(\mathbf{h})$, the set of spherical image intensities of the region of interest, obtained after the transfer of the reference one using the homography, the error to be regulated is:

$$\mathbf{e} = \mathbf{I}_\mathcal{S}(\mathbf{h}) - \mathbf{I}^*_\mathcal{S}. \quad (7)$$

Imposing an exponential decrease of this error, we have $\dot{\mathbf{e}} = -\lambda\mathbf{e}$. The time variation of the intensity at a given point on the sphere is de ned thanks to the Cartesian spherical optical ow constraint equation (OFCE). Assuming a temporal luminance consistency for a moving spherical image point:

$$\mathbf{I}_\mathcal{S}(\mathbf{X}_\mathcal{S} + \Delta\mathbf{X}_\mathcal{S}, t + \Delta t) = \mathbf{I}_\mathcal{S}(\mathbf{X}_\mathcal{S}, t). \quad (8)$$

and if $\Delta\mathbf{X}_\mathcal{S}$ is small enough, the OFCE is valid:

$$\nabla\mathbf{I}_\mathcal{S}^\mathsf{T}\dot{\mathbf{X}}_\mathcal{S} + \dot{\mathbf{I}}_\mathcal{S} = 0. \quad (9)$$

$\nabla\mathbf{I}_\mathcal{S}$ is the spatial Cartesian spherical gradient of $\mathbf{I}_\mathcal{S}(\mathbf{X}_\mathcal{S}, t)$. Hence, the time variation of $\mathbf{I}_\mathcal{S}$ is then expressed as:

$$\dot{\mathbf{I}}_\mathcal{S} = -\nabla\mathbf{I}_\mathcal{S}^\mathsf{T}\dot{\mathbf{X}}_\mathcal{S}. \quad (10)$$

$\dot{\mathbf{X}}_\mathcal{S}$ can be decomposed as:

$$\dot{\mathbf{X}}_\mathcal{S} = \frac{\mathrm{d}\mathbf{X}_\mathcal{S}}{\mathrm{d}t} = \frac{\partial\mathbf{X}_\mathcal{S}}{\partial\mathbf{h}}\frac{\mathrm{d}\mathbf{h}}{\mathrm{d}t} = \mathbf{G}_{\mathbf{X}_\mathcal{S}\mathbf{h}}\dot{\mathbf{h}} = \left[\mathbf{J}_{\mathbf{X}_\mathcal{S}} \mid \mathbf{L}_{\mathbf{X}_\mathcal{S}}\right]\begin{bmatrix}\dot{\mathbf{n}}\\ \dot{d}\\ \mathbf{v}\end{bmatrix} \quad (11)$$

where $\mathbf{v} = (\boldsymbol{\upsilon}, \boldsymbol{\omega})^\mathsf{T}$ is the camera relative pose velocity. Jacobian $\mathbf{J}_{\mathbf{X}_\mathcal{S}}$ expresses the variation of a spherical point w.r.t. the 3D plane (constrained normal and distance), and

$\mathbf{L}_{\mathbf{X}_\mathcal{S}}$ is known as the interaction matrix (for a relative pose, here) and is [10]:

$$\mathbf{L}_{\mathbf{X}_\mathcal{S}} = \left(\tfrac{1}{\rho}(\mathbf{X}_\mathcal{S}\mathbf{X}_\mathcal{S}^T - \mathbf{I}_{3\times3}) \qquad [\mathbf{X}_\mathcal{S}]_\times\right). \tag{12}$$

$\mathbf{L}_{\mathbf{X}_\mathcal{S}}$ mainly depends on the coordinates of a spherical point, apart from ρ, which is the distance of the 3D point in the current camera frame (eq. 1). This distance is obtained by intersecting a viewline and the 3D plane.

Equations (10) and (11) lead to the image Jacobian related to $\mathbf{I}_\mathcal{S}$ at spherical point $\mathbf{X}_\mathcal{S}$:

$$\mathbf{G}_{\mathbf{I}_\mathcal{S}}(\mathbf{X}_\mathcal{S}) = -\nabla\mathbf{I}_\mathcal{S}^\mathsf{T}\mathbf{G}_{\mathbf{X}_\mathcal{S}\mathbf{h}}. \tag{13}$$

D. Tracking of several planes

To deal with a moving camera and tracking multiple planes, the problem can be formulated to compute only one relative pose and several normal and distances. In such a case, intensities of a planar region are used to compute their corresponding plane whereas the intensities of all the planar regions are used to compute the relative pose. Hence, the Jacobian linking the variation of image intensities to the camera motion and parameters of p planes is:

$$\mathbf{G}_{\mathbf{I}_\mathcal{S}} = \begin{bmatrix} -\nabla\mathbf{I}_\mathcal{S}^{(1)} \\ -\nabla\mathbf{I}_\mathcal{S}^{(2)} \\ \dots \\ -\nabla\mathbf{I}_\mathcal{S}^{(p)} \end{bmatrix}^\mathsf{T} \left[\begin{array}{cccc|c} \mathbf{J}_{\mathbf{X}_\mathcal{S}}^{(1)} & 0 & \dots & 0 & \mathbf{L}_{\mathbf{X}_\mathcal{S}}^{(1)} \\ 0 & \mathbf{J}_{\mathbf{X}_\mathcal{S}}^{(2)} & \dots & \vdots & \mathbf{L}_{\mathbf{X}_\mathcal{S}}^{(2)} \\ \vdots & \dots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & \mathbf{J}_{\mathbf{X}_\mathcal{S}}^{(p)} & \mathbf{L}_{\mathbf{X}_\mathcal{S}}^{(p)} \end{array}\right] \tag{14}$$

where (i) shows the plane number to which the Jacobian is related to and its corresponding feature set. Then, the error vectors corresponding to each plane are stacked together and are related, thanks to the latter multiple planes Jacobian, to the camera and planes parameters velocity:

$$\mathbf{V} = \begin{bmatrix}\dot{\mathbf{n}}^{(1)} & \dot{d}^{(1)} & \dot{\mathbf{n}}^{(2)} & \dot{d}^{(2)} & \dots & \dot{\mathbf{n}}^{(p)} & \dot{d}^{(p)} & \mathbf{v}\end{bmatrix}^\mathsf{T} \tag{15}$$

E. Parameters update

The time variation of planes and relative pose vector is computed thanks to a Levenberg-Marquart (LM) optimization method:

$$\mathbf{V} = -\lambda\left(\mathbf{Q} + \mu\,\mathrm{diag}(\mathbf{Q})\right)^{-1}\mathbf{G}_{\mathbf{I}_\mathcal{S}}^\mathsf{T}\begin{bmatrix}\mathbf{I}_\mathcal{S}(\mathbf{h}^{(1)}) - \mathbf{I}_\mathcal{S}^* \\ \vdots \\ \mathbf{I}_\mathcal{S}(\mathbf{h}^{(p)}) - \mathbf{I}_\mathcal{S}^*\end{bmatrix} \tag{16}$$

with $\mathbf{Q} = \mathbf{G}_{\mathbf{I}_\mathcal{S}}^\mathsf{T}\mathbf{G}_{\mathbf{I}_\mathcal{S}}$ and $\mathbf{h}^{(i)} = \left(\mathbf{n}^{(i)}, \mathbf{d}^{(i)}, \mathbf{r}\right)^\mathsf{T}$. If μ is very high, the optimization behaves like a steepest descent whereas a very low value for μ leads equation (16) to behave like a Gauss-Newton method. LM is often used since it ensures a better convergence, thanks to a controlled and ef cient step.

The relative pose is updated at iteration it with the exponential map of $se(3)$ [13] using ${}^{c_2}\mathbf{M}_{c_1}^{it} = {}^{c_2}\mathbf{M}_{c_1}^{it-1}e^{[\mathbf{v}]}$. Plane parameters are updated by addition of their time variation.

F. Gradients computation on the sphere

Computing $\nabla\mathbf{I}_\mathcal{S}$ means to compute the spherical image derivation along $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ axes. Since the spherical image is a surface which can be minimally parameterized using azimuth and elevation angles, ϕ and θ, spherical image gradients are computed w.r.t. to these parameters and then transformed to get the 3D image gradient.

Omnidirectional image ltering is only valid on the equivalent sphere where convolution is commutative with regular sampling. The technique to compute gradients is to use intensities from a spherical neighborhood for the convolution [6].

The two linear neighborhoods $\phi_\mathcal{N}$ and $\theta_\mathcal{N}$ of length l, considering $\Delta_{\phi,\theta} = \arccos(|\begin{pmatrix}0 & 0 & 1\end{pmatrix}.pr_\xi^{-1}(\begin{bmatrix}u_0+1 & v_0 & 1\end{bmatrix}^\mathsf{T})|)$, are expressed as:

$$\begin{cases}\phi_\mathcal{N} = \left\{\left(\theta,\ \phi + k\Delta_\phi\right)^\mathsf{T}, -\frac{l}{2} \le k \le \frac{l}{2}, k \neq 0\right\} \\ \theta_\mathcal{N} = \left\{\left(\theta + k\Delta_\theta,\ \phi\right)^\mathsf{T}, -\frac{l}{2} \le k \le \frac{l}{2}, k \neq 0\right\}\end{cases}. \tag{17}$$

Neighborhoods are then projected on the image plane to retrieve intensities. Nearest neighbor interpolation is done to make the process fast and to allow precomputation of spherical neighbors projected in the image plane for each pixel. These neighborhoods are used with an image derivative lter.

Finally, gradient values I_ϕ, I_θ are used to compute $I_{X_\mathcal{S}}$, $I_{Y_\mathcal{S}}$ and $I_{Z_\mathcal{S}}$, where $I_A = \frac{\partial I}{\partial A}$ knowing the Cartesian-polar coordinates transformation, $\phi = \arccos(Z_\mathcal{S})$ and $\theta = \arctan(Y_\mathcal{S}/X_\mathcal{S})$, and using partial derivatives $\frac{\partial\phi,\theta}{\partial\mathbf{X}_\mathcal{S}}$.

III. Stereoscopic spherical 3D plane tracking

Relative pose estimation in a monocular plane tracking, even with several planes, is estimated up to a scale factor. Dealing with a calibrated stereo rig allows to x the scale and to have planes and relative position in meters.

Figure 2 shows the geometric relations between corresponding points in spherical views. Based on these considerations, the stereoscopic plane tracking has two main stages. First, the search for regions in the three views corresponding to the reference template is made, estimating the 3D plane parameters. Then, for any new stereo image, the relative pose is computed simultaneously to the plane parameters optimization. Therefore, we estimate only a minimal set of parameters.

A. Stereoscopic model

For a stereo rig of N cameras, each camera is modeled using the spherical model. Relative poses of $N-1$ cameras c_j, $j = 1..N$, of the rig, modeled by homogeneous matrices ${}^{c_j}\mathbf{M}_{c_1}$, are de ned w.r.t. the reference one, c_1. In the present case, since the used stereo sensor for our experiments is composed of one orthographic camera and four parabolic mirrors, we consider a stereo rig of four central cameras ($N = 4$). Simultaneous intrinsic and extrinsic calibration has been done using an extension of a stereo omnidirectional virtual visual servoing approach [5].

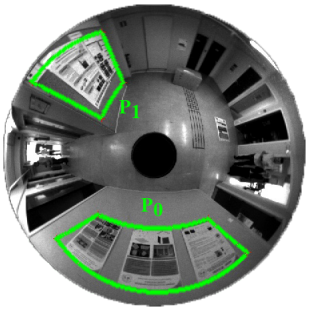
[12] B.D. Lucas and T. Kanade, *An Iterative Image Registration Technique with an Application to Stereo Vision*, Int. Joint Conf. on Arti cial Intelligence, 1981.

[13] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry, *An invitation to 3-D vision*, Springer, 2004.

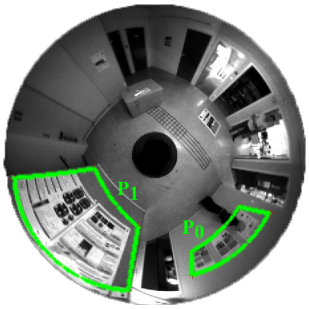
[14] E. Marchand and F. Chaumette, *Fitting 3D Models on Central Catadioptric Images*, IEEE Int. Conf. on Robotics and Automation (Roma, Italia), April 2007.

[15] C. Mei, S. Benhimane, E. Malis, and P. Rives, *Ef cient Homography-Based Tracking and 3-D Reconstruction for Single-Viewpoint Sensors*, IEEE Trans. on Robotics 24 (2008), no. 6.

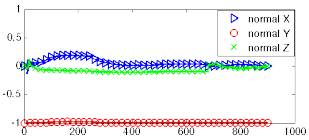
[16] A. Salazar-Garibay, E. Malis, and C. Mei, *Visual Tracking of Planes with an Uncalibrated Central Catadioptric Camera*, IEEE/RSJ Int. Conf. on Intelligent RObots and Systems (Saint-Louis, USA), October 2009.



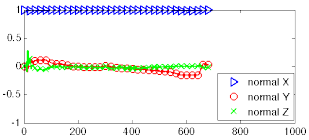
(a) Image 0 / 900



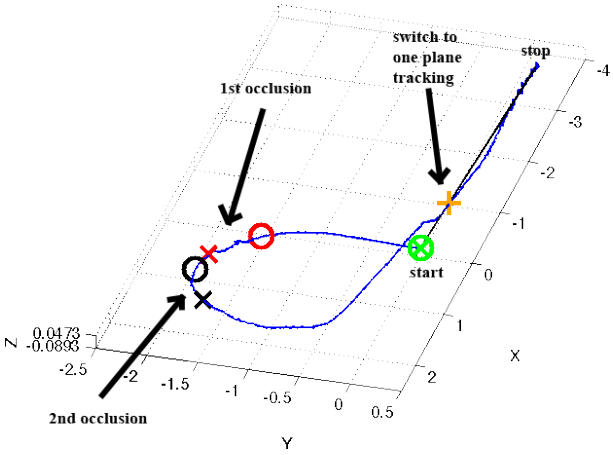
(b) Image 216 / 900 ($\mathbf{P}_0$ occluded)



(c) Evolution of the $\mathbf{P}_0$ normal.

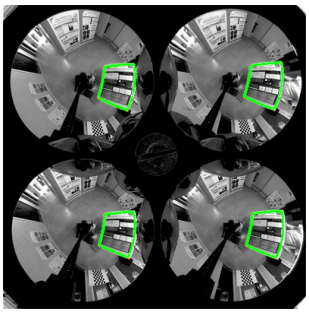


(d) Evolution of the $\mathbf{P}_1$ normal.

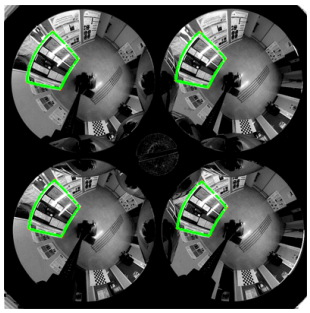


(e) Estimated trajectory with a circled green cross marking the starting position and two sets of a circle and a cross showing the trajectory parts when tracked planes are occluded (red for $\mathbf{P}_0$ and black for $\mathbf{P}_1$). The orange cross marks the end of the tracking of $\mathbf{P}_1$ and a straight black line links the start position to the end.

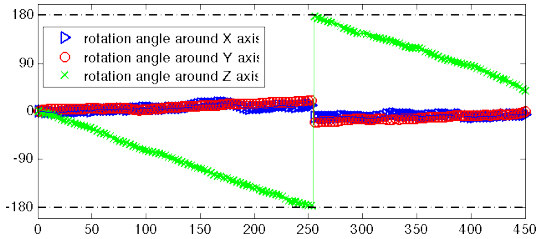
Fig. 3. Curve motion along 11 m with partial occlusion. Planes $\mathbf{P}_0$ and $\mathbf{P}_1$ are simultaneously tracked in a sequence of 900 images until $\mathbf{P}_1$ is no more visible. The rest of the estimation is done with the tracking of $\mathbf{P}_1$ (from the orange cross on trajectory (e)). To have the trajectory at the good scale in meter, the distance to $\mathbf{P}_0$ is manually measured from the initial camera position. The nal position error is below 1 cm (unmeasurable) and the standard deviation of the estimation of the Z coordinate is about 2.64 cm for the complete estimated trajectory.



(a) Image 0 / 450

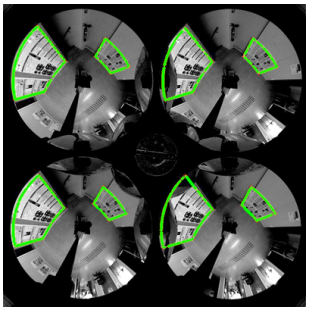


(b) Image 220 / 450

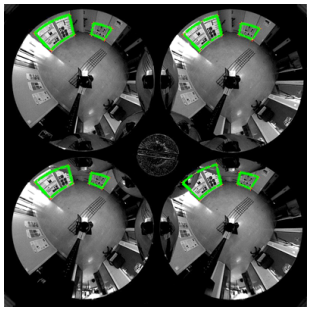


(c) Estimated angles of rotation w.r.t. the reference image.

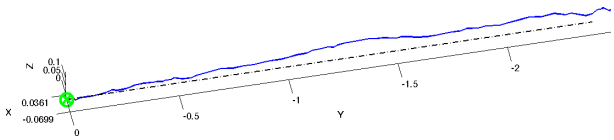
Fig. 4. Stereoscopic tracking of a plane all around the sensor. The rotation of the planar object is nearly pure around the camera axis (c).



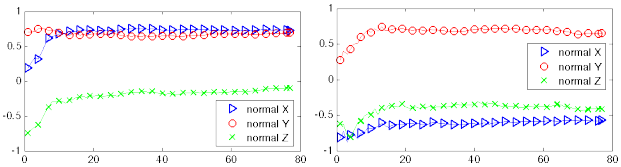
(a) Image 0 / 230



(b) Image 230 / 230



(c) Estimated trajectory in blue and reference one in dashed black. The unit is the meter and the circled green cross shows the start position.



(d) Plane normal vectors evolution along the sequence.

Fig. 5. Straight motion along 2.40 m. The nal position error is 10 cm in 3D leading to an error ratio over the path length of around 4%. Note that the smallest plane in images (a) and (b) is not vertical, that is why its Z component is not 0.